DigItalia

Anno V, Numero 2 - **2010**

Rivista del digitale nei beni culturali

ICCU-ROMA

Delivering Content to Europeana in Practice: The ATHENA Harvesting Format LIDO*

Regine Stein
Philipps-Universität Marburg

Introduction

ATHENA (Access to cultural heritage networks across Europe) provides content to Europeana by establishing a mechanism for harvesting museum holdings into Europeana. A major goal of the project is to develop an infrastructure that enables semantic interoperability with Europeana while preserving museum object specifics. To comply with this requirement, ATHENA has put particular focus on the choice of a format for content delivery that would be able to express the variety of museum's information appropriately. While the practical harvesting of data is an ongoing process and experience is growing, this article provides together with the actual presentation of the ATHENA harvesting format LIDO, some preliminary conclusions derived from the project.

ATHENA's choice of a data model: ESE versus LIDO

The data model currently used in the Europeana prototype, ESE, is based on the Dublin Core metadata format. Although initially created strictly for the description of Web resources, Dublin Core has become the most common format in cultural heritage service environments. However, the ESE model is not considered as appropriate within the museum community: museum metadata is "flatten out", with most of the data going into a limited subset of elements. For example, a number of different persons and institutions are usually associated with a museum object: the creator or finder of an object, important persons who have used it, the museum currently holding it, previous owners, and so on. All this qualified information is lost in the ESE format. Moreover, the lack of structure that allows elements to be grouped according to their semantic content leads to substantial information loss. A particular problem is the fact that Dublin Core does not allow information about the object itself and its digital surrogate to be clearly differentiated – the creator of the object appears in the same field than the photographer of its image.

Consequently, the ATHENA workpackage on metadata formats, following a best practice report on metadata formats used by the partners, came to the conclusion that a more appropriate data model for museum information should be used. Since the LIDO development already underway was primarily an effort to harmonize the two existing harvesting formats CDWA Lite and museumdat into one single schema, ATHENA decided to join the LIDO initiative and support further development that would subsequently integrate SPECTRUM requirements into the schema. Thus LIDO was

* This article is a shortened version of the full paper *Sharing Museum Information: Theory or Practice – A European Experience* given at the CIDOC 2010 conference in Shanghai, November 10th 2010. For the full paper refer to http://cidoc.meta.se/2010/full_papers/stein.pdf.

chosen and further developed as the metadata format for the delivery of museum content through ATHENA to Europeana.

The LIDO format

LIDO is an XML schema intended for delivering metadata, for use in a variety of online services, from an organization's online collections database to portals of aggregated resources, as well as exposing, sharing and connecting data on the Web. The strength of LIDO lies in its ability to support the full range of descriptive information about museum objects; it can be used for all kinds of object, e.g. art, cultural, technology and natural science. Moreover, it supports multilingual portal environments.

LIDO defines 14 groups of information of which just three are mandatory. This allows for the widest and most comprehensive range of information possible. Organizations can decide on how rich – or how light – they want their contributed metadata records to be.

The schema consists of a nested set of "wrapper" and "set" elements, many of them repeatable, which organizes information about an object into a tree-like structure. This allows any degree of detail to be recorded in a logically correct, semantically coherent way. An important part of its design is the concept of events, taken from the CIDOC CRM. Information about actors, dates and places related to a museum object is mediated through an event: the creation, collection, and use of an object are seen as events occurring during the object's lifecycle. An exception is events that are depicted or referred to directly, considered as subject matter.

Another important construction principle is the distinction between indexing information that is optimized for searching and retrieval, and display information that is optimized for online presentation. Each information unit contains distinct sub-elements for indexing and display.

The structural elements of LIDO contain "data elements" which hold actual data *values*. LIDO also allows the recording of information about data *sources* (e.g. in a book) and references to controlled terminology (e.g. the identification code for a term in a thesaurus). Conceptually the information in a LIDO record is organized in 7 areas, of which 4 have descriptive and 3 an administrative character:

The descriptive information section holds:

- object classification information such as object type and other classifications;
- object identification information such as titles, inscriptions, repository information, descriptions, and measurements;
- event information about events where the object was present or in which it participated, such as creation, modification, acquisi-

Figure 1. LIDO overview

tion, finding, or use. This section holds a number of sub-elements including event type and name, participating actors, cultures involved, date and place information as well as materials and techniques used (typically in the creation/production event);
– relation information links to related objects, but also to the subject – that is the content of a work: what is depicted in or by a work or what the work is about.

The administrative information section holds:

– rights associated with the object;
– record information about the source providing the metadata;
– resource information, in particular about digital resources being supplied to the service environment for representing an object online.

The result of a joint effort of several international key institutions and groups dealing with museum documentation standards, e.g. the CDWA, museumdat, SPECTRUM and CIDOC CRM communities, the release of LIDO v. 1.0 during this year's CIDOC conference can be seen as a clear reward to the community. It provides a single, common schema for contributing content to cultural heritage repositories. This enables museums and other content providers, using different data structures and software systems, to express and deliver a wide variety of information in a standardized and machine-readable format. Furthermore, this information can easily be accessed, harvested and recontextualized by semantic-aware services. Apart from the exciting promise of new applications, LIDO promises time – and cost – savings for museums interchanging object information in different daily work contexts.

The ATHENA mapping and ingestion process

Now after this insight into the richness and opportunities of LIDO, the question arises as to how manageable the mapping and ingestion process is for content providers who may have only recently started sharing their data in a wider service environment. To facilitate this process a mapping tool has been developed by the technical partner of the ATHENA project, the National Technical University of Athens.
Any kind of data provided in an XML format can be loaded into the system. The tool then visualizes, on the left, the incoming source data structure and, on the right, the LIDO target schema. The content provider can then map its source data fields through drag and drop to the target fields, including mapping of structural elements holding no data, and conditions for the mapping and concatenation of data values and constants. A helpdesk mailing list allows users to ask questions about the format and the tool, and to help each other.
Combining a comprehensive metadata format with a customized technical solution for practical mapping is an exciting effort. It enables semantic interoperability of content from many different collections and from different management systems with different data structures. It is difficult to evaluate how the process will evolve over the next few months of the ATHENA project's activities and beyond, but some preliminary statements may be given here for discussion, both, positive and instructive. The overall mapping results are good and the questions on the helpdesk list comprehensive, so users appear to have grasped, from the material and the tool provided, both the LIDO schema and how to map to it.
Yet to get to a full and meaningful mapping that best reflects the source information in the target schema, several feedback loops are often needed between the local expert, who knows the source schema and content very well, and a LIDO expert who knows the LIDO structure in depth. This loop is considerably shortened by the ATHENA mapping tool, the result of a close cooperation between LIDO schema developers and technical imple-

Figure 2. ATHENA mapping tool

menters, which reflects the target schema very clearly. The process is considerably easier if the source schema is based on a documentation standard such as SPECTRUM or national standard. Moreover, features supporting data analysis and data value statistics, such as provided in the mapping tool, help immensely in this process.

Conclusion

Overall it seems that it is both appropriate and simpler for content providers to map their data to a well-structured metadata format, instead of randomly choosing some corresponding field in a flat structure such as ESE.

Presently, LIDO serves in ATHENA as an intermediate layer between source formats and the Dublin Core-based ESE format. It thereby provides a more standardized representation of museum collections in Europeana. Since the ESE format does not support the fine granularity of museum information and fails to make a clear distinction between the museum object itself and its digital surrogate in an online service, standardized presentation helps to improve search and display quality considerably.

It will be crucial to see now the practical implementation of the new Europeana Data Model, EDM. EDM will supplement and enhance the currently used ESE model with a meta-structure that truly allows the LIDO format to be retrieved. It is a clear expectation that the implementation of this data model will significantly improve resource discovery, providing more precise search results that carry meaningful links to associated resources.

LIDO effectively prepares the ground for such new, data quality focused approaches. Used in conjunction with increasing opportunities to participate in linked data environments – as they are aimed at in the forthcoming EU-funded Linked Heritage project, this will enable museums to recontextualize their collections in a meaningful way and hence improve understanding of the collections within the greater cultural heritage context.

For full reference of LIDO visit http://www.lido-schema.org/.

Several training material can be found at http://www.athenaeurope.org/index.php?en/159/training.